FINAL EXAM - CSC/DSC 262/462 - December 20, 2017

NAME: ____

PLEASE NOTE:

Undergraduate students do questions 1-10.

Graduate students do questions 3-12.

The exam will last three hours. Answer each required question in the space provided. All questions have equal weight. Use the back of the page for additional space if needed (please indicate clearly where you have done this). A methodological summary will be provided separately. You may use a calculator, but no aid sheet is permitted. Critical values for the standard normal, t and χ^2 distributions are given in tabular form in the methodological summary (page 8). Critical values for the F distribution are given as needed within specific questions.

Q1 [Undergraduate Students Only]: For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 138.6$, $n = 15$, sample standard deviation $S = 12.04$.

(a) Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.
(b) Calculate a level $1 - \alpha = 0.95$ upper confidence bound for σ.
(c) Using the upper bound for σ calculated in part (b) estimate the sample size needed to obtain a level $1 - \alpha = 0.95$ confidence interval for μ with a margin of error of 3.0. Use a normal approximation.

SOLUTION

(a)

$$
\begin{aligned}
CI_{1-\alpha} &= \bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \\
&= 138.6 \pm 2.14 \times \frac{12.04}{\sqrt{15}} \\
&= 138.6 \pm 6.67 = (131.93, 145.27)
\end{aligned}
$$

(b)

$$
\begin{aligned}
UB &= \frac{S}{\sqrt{\chi^2_{n-1,1-\alpha}/(n-1)}} \\
&= \frac{12.04}{\sqrt{6.57/14}} \\
&= 17.57
\end{aligned}
$$

So,

$$\sigma < 17.57$$

is the 95% upper confidence bound for σ.

(c) Use estimate $\hat{\sigma} = 5.836$ in formula

$$n \approx \left(z_{\alpha/2} \frac{\hat{\sigma}}{E_o} \right)^2 = \left(1.96 \times \frac{17.57}{3} \right)^2 = 130.43,$$

so round up to $n = 131$.

Q2 [Undergraduate Students Only]: We are given two independent samples from normally distributed populations $N(\mu_i, \sigma_i^2)$, $i = 1, 2$. The data is summarized in the table below.

(a) Perform a two-sided hypothesis test for null hypothesis $H_o : \sigma_1^2 = \sigma_2^2$ against alternative $H_a : \sigma_1^2 \neq \sigma_2^2$. Use significance level $\alpha = 0.05$. You can make use of critical values $F_{0.975,22,53} = 0.463$ and $F_{0.025,22,53} = 1.943$.

(b) Construct a level $1 - \alpha = 0.9$ confidence interval for $\mu_2 - \mu_1$. Use the conclusion of part (a) to choose between the pooled procedure for equal variances or Welch's procedure for unequal variances.

	Sample $i = 1$	Sample $i = 2$
$\bar{X}_i$	43.96	48.98
S_i	7.62	9.42
n_i	23	54

SOLUTION

(a) Use statistic

$$F = \frac{S_1^2}{S_2^2} = \frac{7.62^2}{9.42^2} = 0.654.$$

Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F \leq F_{1-\alpha/2,n_1-1,n_2-1} = 0.463 \text{ or } F \geq F_{\alpha/2,n_1-1,n_2-1} = 1.943.$$

Therefore, do not reject the null hypothesis of equal variances at a signficance level $\alpha = 0.05$.

(b) Use the pooled procedure with $\nu = n_1 + n_2 - 2 = 75$ degrees of freedom. Pooled variance is given by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{22 \times 7.62^2 + 53 \times 9.42^2}{75} = 79.74.$$

The confidence interval is

$$\begin{aligned} CI_{1-\alpha} &= \bar{X}_2 - \bar{X}_1 \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 48.98 - 43.96 \pm 1.665 \times 8.93 \sqrt{\frac{1}{23} + \frac{1}{54}} \\ &= 5.02 \pm 3.70 \\ &= (1.32, 8.72). \end{aligned}$$

Q3 [All Students]: We are given two paired samples from normally distributed populations ($n = 5$). The data is summarized in the table below. Perform a two-sided hypothesis test using hypotheses $H_o : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$. Use significance level $\alpha = 0.1$.

	Sample 1	Sample 2	Difference
1	10.94	9.49	1.45
2	13.74	12.86	0.88
3	14.27	11.10	3.17
4	14.30	10.84	3.46
5	13.58	10.09	3.49

SOLUTION From the table we have:

$$\bar{X}_1 = 13.365,\ \bar{X}_2 = 10.875,\ \bar{X}_2 - \bar{X}_1 = 2.49,\ S_D = 1.233.$$

Test statistic is

$$\begin{aligned} T &= \frac{\bar{D}}{S_D/\sqrt{n}} \\ &= \frac{\bar{X}_2 - \bar{X}_1}{S_D/\sqrt{n}} \\ &= \frac{2.49}{1.233/\sqrt{5}} \\ &= 4.517. \end{aligned}$$

Reject H_o if

$$|T| \geq t_{n-1,\alpha/2} = t_{4,0.05} = 2.132.$$

Therefore, reject the null hypothesis at a signficance level $\alpha = 0.1$.

Q4 [All Students]: Suppose a binomial random variable $X \sim bin(n, p)$ is observed to be $X = 21$, with sample size $n = 80$.

(a) Construct a level $1 - \alpha = 0.95$ confidence interval for p. Use the normal approximation.

(b) Test hypothesis $H_o : p \geq 0.4$ against $H_a : p < 0.4$. Is the null hypothesis rejected at a significance level of $\alpha = 0.05$? Use a normal approximation with a continuity correction

SOLUTION

(a) The level $1 - \alpha$ confidence interval for p is given by

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The estimate of p, given $X = 21$ and $n = 80$ is

$$\hat{p} = \frac{X}{n} = \frac{21}{80} = 0.262.$$

The confidence interval is then given by

$$\begin{aligned} CI &= 0.262 \pm 1.96\sqrt{\frac{0.262(1-0.262)}{80}} \\ &= 0.262 \pm 1.96 \times 0.0492 \\ &= 0.262 \pm 0.0964 \end{aligned}$$

or equivalently, $CI = (0.166, 0.359)$.

(b) To implement the continuity correction, first express statistic Z_{obs} in terms of the counts:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{X - np_0}{\sqrt{np_0(1-p_0)}}.$$

Note that np_0 need not be an integer. Then the continuity correction can be implemented by using the corrected statistic

$$Z'_{obs} = \frac{X - np_0 + 0.5}{\sqrt{np_0(1-p_0)}} = \frac{21 - 32 + 0.5}{\sqrt{80 \times 0.4(1-0.4)}} = -2.396.$$

The critical value is $-z_{.05} = -1.645$. Since $Z'_{obs} < -z_{.05}$ we reject the null hypothesis at an $\alpha = 0.05$ significance level.

Q5 [All Students]: We are given an *iid* sample from a normal distribution

$$91.84, 103.35, 94.56, 95.42, 105.72,$$

of sample size $n = 5$. Calculate a confidence interval for population standard deviation σ, using confidence level $1 - \alpha = 0.95$. Also give the level $1 - \alpha = 0.95$ lower confidence bound.

SOLUTION The sample standard deviation is $S = 6.01$. The level $1 - \alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{4,0.025} = 11.143 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{4,0.975} = 0.484.$$

The confidence interval is then given by

$$\frac{6.01}{\sqrt{11.143/4}} < \sigma < \frac{6.01}{\sqrt{0.484/4}}$$

or equivalently, $CI = (3.601, 17.271)$.

The level $1 - \alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{4,0.05} = 9.488$. The lower bound is then given by,

$$\sigma > \frac{6.01}{\sqrt{9.488/4}} = 3.903.$$

Q6 [All Students]: Suppose we are given the following contingency table, summarizing the infection history of $n = 1621$ subjects.

	Not Vaccinated	Vaccinated	
Infection occurs	34	7	41
No infection occurs	933	647	1580
Total	967	654	1621

Construct a level $1 - \alpha = 0.95$ confidence interval for the log odds ratio for infection occurence between groups Not Vaccinated and Vaccinated. Can you reject the null hypothesis $H_o : OR = 1$ against $H_a : OR \neq 1$ at significance level of $\alpha = 0.05$?

SOLUTION

(a) The estimate of the odds ratio is given by

$$OR = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{34 \times 647}{933 \times 7} = 3.368.$$

We use critical value

$$z_{\alpha/2} = z_{0.025} = 1.96.$$

The standard error of the estimate $\log(OR)$ is

$$\begin{aligned} SE(\log(OR)) &= \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \\ &= \sqrt{\frac{1}{34} + \frac{1}{7} + \frac{1}{933} + \frac{1}{647}} \\ &= 0.418. \end{aligned}$$

The level $1 - \alpha$ confidence interval for the odds ratio is given by

$$\begin{aligned} CI &= \log(OR) \pm z_{\alpha/2} SE(\log(OR)) \\ &= 1.214 \pm 1.96 \times 0.418 \\ &= 1.214 \pm 0.82 \\ &= (0.394, 2.034) \end{aligned}$$

or equivalently, $CI = (0.395, 2.034)$. Since the CI does not contain 0 we reject the null hypothesis at an $\alpha = 0.05$ significance level.

Q7 [All Students]: The observed counts for $k = 3$ categories based on a random sample of size $n = 224$ are given in the following table. Hypothetical population frequencies p_i^o are also given in the table. Use a χ^2 goodness of fit test for null and alternative hypotheses:

$$H_o : p_i = p_i^o \text{ for all } i = 1, 2, 3 \text{ against } H_a : p_i \neq p_i^o \text{ for some } i = 1, 2, 3,$$

where $p_i^o = 1/3$, $i = 1, 2, 3$. Use significance level $\alpha = 0.05$. Do not use Yates's correction.

$i =$	1	2	3	Totals
Observed counts O_i	39	80	105	224
Hypothetical frequencies p_i^o	1/3	1/3	1/3	1.00
Observed frequencies $\hat{p}_i$	0.17	0.36	0.47	1.00

SOLUTION The expected counts are

$$E_i = np_i^o = 224 \times 1/3 = 74.67, \quad i = 1, 2, 3,$$

and are given in the table below. There are $k = 3$ categories, so the χ^2 statistic has $k - 1 = 2$ degrees of freedom. The appropriate ciritcal value is $\chi^2_{2,0.05} = 5.991$. The statistic is given by

$$X^2 = \sum_{i=1}^{3} (O_i - E_i)^2/E_i = 17.04 + 0.38 + 12.32 = 29.74.$$

Note that the terms for $i = 1, 3$ each exceed $\chi^2_{2,0.05}$ so only one of these needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

	1	2	3	Totals
Observed counts O_i	39	80	105	224
Expected counts E_i	74.67	74.67	74.67	224.00
$(O_i - E_i)^2/E_i$	17.04	0.38	12.32	29.74

Q8 [All Students]: A contingency table with $n_r = 2$ rows and $n_c = 3$ columns based on a random sample of size $n = 231$ is given below. Hypothetical population frequencies of cell i, j are given by $p_{i,j}$. The population frequencies for the marginal row i and column j categories are given by r_i and c_j, respectively. Use a χ^2 test for the null hypothesis of row and column independence $H_o : p_{i,j} = r_i c_j$ for all i, j. Use significance level $\alpha = 0.05$. Do not use Yates's correction.

	1	2	3	Totals
1	134	23	37	194
2	7	17	13	37
Totals	141	40	50	231

Table 1: Observed counts $O_{i,j}$

SOLUTION The expected counts are

$$E_{i,j} = R_i C_j / n, \quad i = 1, 2; \quad j = 1, 2, 3,$$

where R_i, C_j are the row and column totals and $n = 231$ is the total count. The values are given in the table below. The χ^2 statistic has $(n_r - 1)(n_c - 1) = 1 \times 2 = 2$ degrees of freedom. The appropriate ciritcal value is $\chi^2_{2,0.05} = 5.991$. The statistic is given by

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} (O_{i,j} - E_{i,j})^2 / E_{i,j} = 37.364.$$

Note that the term for $i, j = 2, 1$ and $i, j = 2, 2$ exceeds $\chi^2_{2,0.05}$ so only this one needs to be calculated. Therefore, reject the null hypothesis at a signficance level $\alpha = 0.05$.

	1	2	3	Totals
1	118.42	33.59	41.99	194.00
2	22.58	6.41	8.01	37.00
Totals	141.00	40.00	50.00	231.00

Table 2: Expected counts $E_{i,j}$

	1	2	3	Totals
1	2.05	3.34	0.59	5.98
2	10.75	17.51	3.11	31.38
Totals	12.81	20.85	3.70	37.36

Table 3: X^2 statistic terms $(O_i - E_i)^2 / E_i$

Q9 [All Students]: We are given two paired samples of sample size $n = 7$. The data is summarized in the table below. Suppose $\tilde{\mu}_D$ is the population median of the paired differences $D = X - Y$. Perform a signed rank test using hypotheses $H_o : \tilde{\mu}_D = 0$ against $H_a : \tilde{\mu}_D \neq 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Sign
1	23.9	12.3	11.6	+
2	28.1	31.5	-3.4	-
3	26.7	24.4	2.3	+
4	32.4	29.7	2.7	+
5	20.7	21.9	-1.2	-
6	23.0	26.1	-3.1	-
7	27.8	26.5	1.3	+

SOLUTION

	Sample 1 (X)	Sample 2 (Y)	Difference ($D = X - Y$)	Rank $\|D\|$	Sign
1	23.9	12.3	11.6	7.0	+
2	28.1	31.5	-3.4	6.0	-
3	26.7	24.4	2.3	3.0	+
4	32.4	29.7	2.7	4.0	+
5	20.7	21.9	-1.2	1.0	-
6	23.0	26.1	-3.1	5.0	-
7	27.8	26.5	1.3	2.0	+

There are no ties, so we have $n = 7$ pairs. The negative and positive rank sums are, respectively,

$$T_- = 12 \text{ and } T_+ = 16.$$

Then

$$T_{obs} = \min(T_-, T_+) = \min(12, 16) = 12.$$

The mean and standard deviation of the negative or positive rank sums are

$$\mu_T = \frac{n(n+1)}{4} = 14 \text{ and } \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = 5.916.$$

This gives z-score

$$Z = \frac{T_{obs} - \mu_T}{\sigma_T} = \frac{12 - 14}{5.916} \approx -0.338$$

We reject H_o if $Z \leq -z_{\alpha/2} = -1.96$. So, do not reject H_o with $\alpha = 0.05$ significance level.

Q10 [All Students]: We are given two independent samples of sample sizes $n_1 = 5$, $n_2 = 11$. The data is summarized in the table below. Suppose $\tilde{\mu_i}$ is the population median of sample i. Perform a two-sided rank sum test using hypotheses $H_o : \tilde{\mu_1} - \tilde{\mu_2} = 0$ against $H_a : \tilde{\mu_1} - \tilde{\mu_2} \neq 0$. Use significance level $\alpha = 0.05$. Use a normal approximation without continuity correction.

	1	2	3	4	5	6	7	8	9	10	11
Sample 1	86.2	79.7	79.8	92.8	75.9						
Sample 2	125.9	122.9	120.0	120.3	97.5	122.0	127.3	127.0	101.8	90.1	123.5

SOLUTION

	1	2	3	4	5	6	7	8	9	10	11
Sample 1	86.2	79.7	79.8	92.8	75.9						
Sample 2	125.9	122.9	120.0	120.3	97.5	122.0	127.3	127.0	101.8	90.1	123.5
Ranks 1	4.0	2.0	3.0	6.0	1.0						
Ranks 2	14.0	12.0	9.0	10.0	7.0	11.0	16.0	15.0	8.0	5.0	13.0

The sample medians are $\tilde{X}_1 = 79.8$ and $\tilde{X}_2 = 122$. The rank sum for samples 1 and 2 are, respectively, $T_1 = 16$ and $T_2 = 120$. The mean and standard deviation of T_1 are

$$\mu_1 = n_1(n_1 + n_2 + 1)/2 = 5 \times (5 + 11 + 1)/2 = 42.5$$

and

$$\sigma_W^2 = n_1 n_2 (n_1 + n_2 + 1)/12 = 77.92, \quad \sigma_W = \sqrt{105} = 8.827.$$

This gives Z-score

$$Z = \frac{T_1 - \mu_1}{\sigma_W} = \frac{16 - 42.5}{8.827} = -3.00.$$

Since $Z < -z_{\alpha/2} = 1.96$ we reject H_o for $\alpha = 0.05$.

Q11 [Graduate Students Only]: A conventional treatment for sleep apnea is reported to improve sleep quality within 1 week for 60% of subjects. A new experimental treatment is studied using 12 subjects. A sleep quality index is measured at the start of treatment and after 1 week of treatment for each subject (higher values signify better sleep quality). The results are given in the following table. Is there evidence that more than 60% of subjects experience improved sleep quality within 1 week? Report an exact P-value, and use significance level $\alpha = 0.05$.

Subject	Start of treatment (X)	1 week of treatment (Y)	Difference ($D = Y - X$)
1	29.1	32.4	3.3
2	30.0	39.2	9.2
3	23.3	34.5	11.2
4	26.2	30.8	4.6
5	25.9	39.5	13.6
6	29.7	24.0	-5.7
7	22.8	25.8	3.0
8	23.6	29.3	5.7
9	27.6	36.0	8.4
10	25.1	30.6	5.5
11	31.4	42.4	11.0
12	27.2	33.0	5.8

SOLUTION Let p be the proportion experiencing improved sleep quality. Let X be the observed number of subjects experiencing improved sleep quality. Then $X \sim bin(12, p)$. The appropriate hypotheses are $H_o : p \leq 0.6$ against $H_a : p > 0.6$. The p-value is

$$P(X \geq 11) = \binom{12}{11} \times 0.6^{11} \times 0.4^{1} + \binom{12}{12} \times 0.6^{12} = 12 \times 0.00145 + 0.00218 = 0.0196$$

where $X \sim bin(12, 0.6)$. Since $P < \alpha$ we reject H_o.

Q12 [Graduate Students Only]: Three independent poll samples estimate popular support p for a certain candidate. The estimates and sample sizes are given in the following table.

Poll i	$\hat{p}_i$	n_i
1	0.45	200
2	0.48	1300
3	0.42	750

(a) Suppose we construct a pooled estimate of p by taking the weighted average

$$\hat{p}_{pooled} = \frac{\sum_{i=1}^{3} n_i \hat{p}_i}{\sum_{i=1}^{3} n_i}.$$

Calculate $\hat{p}_{pooled}$ and estimate its standard deviation (use the value of $\hat{p}_{pooled}$ in your estimate).

(b) Suppose we use the following alternative method of constructing a pooled estimate:

$$\hat{p}^*_{pooled} = \frac{\sum_{i=1}^{3} \hat{p}_i}{3}.$$

Calculate $\hat{p}^*_{pooled}$ and estimate its standard deviation (use the value of $\hat{p}^*_{pooled}$ in your estimate).

(c) Which pooled estimator is more accurate?

SOLUTION Note that in each case $\hat{p}_i = X_i/n_i$ where $X_i \sim bin(n_i, p)$.

(a) First note that if $X = \sum_{i=1}^{3} n_i \hat{p}_i$, then $X \sim bin(p, n_1 + n_2 + n_3)$. Then

$$\hat{p}_{pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2 + n_3\hat{p}_3}{n_1 + n_2 + n_3} = \frac{90 + 624 + 315}{200 + 1300 + 750} = \frac{1029}{2250} = 0.4573.$$

Using $\hat{p}_{pooled}$ to estimate p, we have standard deviation

$$\sigma_{\hat{p}} \approx \sqrt{\frac{\hat{p}_{pooled}(1 - \hat{p}_{pooled})}{n_1 + n_2 + n_3}} = \sqrt{\frac{0.4573(1 - 0.4573)}{2250}} = 0.010488.$$

(b) Suppose we use the following alternative method of constructing a pooled estimate:

$$\hat{p}^*_{pooled} = \frac{\sum_{i=1}^{3} \hat{p}_i}{3} = \frac{0.45 + 0.48 + 0.42}{3} = 0.45.$$

Since the estimates are independent,

$$\sigma^2_{\hat{p}} = \frac{1}{3^2}\sigma^2_{\hat{p}_1} + \frac{1}{3^2}\sigma^2_{\hat{p}_2} + \frac{1}{3^2}\sigma^2_{\hat{p}_3}.$$

Using estimate $p \approx \hat{p}^*_{pooled}$ we have

$$\begin{aligned}
\sigma^2_{\hat{p}^*} &= \frac{1}{3^2}\hat{p}^*_{pooled}(1 - \hat{p}^*_{pooled})\left[\frac{1}{n_1} + \frac{1}{n_2} + \frac{1}{n_3}\right] \\
&= \frac{1}{3^2}0.45(1 - 0.45)\left[\frac{1}{200} + \frac{1}{1300} + \frac{1}{750}\right] \\
&= 0.000195.
\end{aligned}$$

This gives

$$\sigma_{\hat{p}^*} \approx \sqrt{0.000195} = 0.01396.$$

(c) Since $0.01396 = \sigma_{\hat{p}^*} > \sigma_{\hat{p}} = 0.010488$, the first pooled estimator $\hat{p}_{pooled}$ is more accurate.